GROUNDINGV2X: GROUNDED PRE-TRAINING FOR MODEL-, TASK-, AND ALIGNMENT-AGNOSTIC COLLABORATIVE PERCEPTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Collaborative perception aims to enhance the perception capability of a single vehicle by enabling information sharing through Vehicle-to-Everything (V2X) communication. However, the heterogeneity of agents poses a significant challenge to the scalability of collaborative perception systems. Existing approaches rely on encoder retraining, feature interpretation, or protocol adaptation to align heterogeneous representations, which are difficult to implement in real-world practice. To address this, we introduce GroundingV2X, a model-, task-, and alignment-agnostic collaborative perception framework via grounded pre-training. Specifically, agents communicate through grounded text derived from their perception results and composed of positional, geometric, and semantic key-value pairs with numerical attributes, providing a unified interface for heterogeneous information exchange. Moreover, we incorporate Fourier numerical embeddings and subsentence level attention to preserve numerical precision and entity-specific context with a lightweight text encoder. Then, a position-aware text-to-feature enhancer selectively integrates the grounded text into spatially relevant local features for collaborative perception. In addition, each agent is pre-trained independently using its perceived grounded text, synthetic grounded text, and annotation-derived pseudo-labels for supervision, improving its generalization to diverse textual descriptions. Extensive experiments on OPV2V, DAIR-V2X, and V2V4Real demonstrate that GroundingV2X enables efficient heterogeneous collaboration solely through pre-training while achieving superior performance compared to state-of-the-art methods both in single-task and multi-task scenarios with substantially lower communication bandwidth requirements.

1 INTRODUCTION

Enabling vehicles to share complementary information with other vehicles or infrastructure through V2X communication, collaborative perception has emerged as a promising solution to address the limitations of single-vehicle perception, such as occlusions and limited sensing range (Han et al., 2023). In the past few years, collaborative perception has substantially advanced the performance over single-vehicle systems, with intermediate feature-level fusion in particular demonstrating distinct advantages in both accuracy and robustness (Wang et al., 2020; Xu et al., 2022b; 2023a). However, these gains rely on well-aligned feature representations across collaborators, while the heterogeneity of agents, such as different sensor equipment, model architectures, and perception tasks, poses a major barrier to the widespread deployment in the real world (Bai et al., 2024).

Recently, several studies have explored solutions for heterogeneous collaboration by aligning the intermediate representations across agents through encoder retraining (Lu et al., 2024), feature interpretation (Luo et al., 2024), or a predefined shared protocol (Gao et al., 2025b). While effective, as illustrated in Figure 1, these approaches require cross-agent feature alignment, collaborative post-training or specific protocol adaptation, which are challenging to implement in real-world scenarios.

To address these challenges, we propose GroundingV2X, a model-, task-, and alignment-agnostic framework that establishes grounded text as the communication interface between heterogeneous agents without either collaborative post-training or cross-agent feature alignment. Specifically, each

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

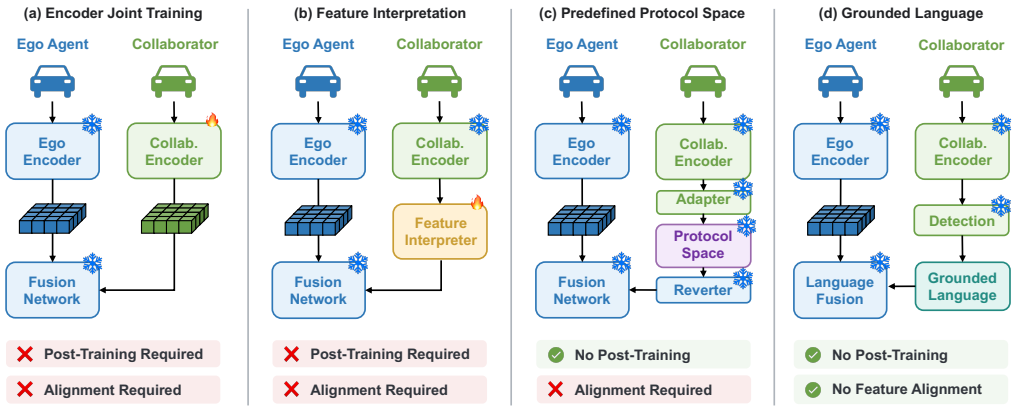


Figure 1: **Comparison of heterogeneous collaboration strategies.** (a) Encoder joint training aligns agents' features. (b) Feature interpretation translates collaborator features into the ego space. (c) Protocol adaptation maps features to a shared protocol. (d) GroundingV2X exchanges grounded text without collaborative post-training or cross-agent feature alignment.

agent constructs grounded text from its perceptual results, which is composed of a set of attribute keywords paired with numerical values, covering the positional, geometric, and semantic information of the detected objects or regions. Upon receiving the grounded texts from other agents, the receiver agent encodes the grounded texts with a lightweight text encoder equipped with Fourier numerical embedding and sub-sentence level attention to capture fine-grained numerical information and extract the context among key-value pairs, avoiding the computational overhead of large-scale language models while enabling precise text understanding. The encoded text features are then selectively integrated into the spatially relevant local features of the receiver through a position-aware text-to-feature enhancer. Furthermore, we generate diverse synthetic grounded texts and pseudo-labels derived from dataset annotations for grounded pre-training, allowing the agent to accommodate varied textual expressions during collaboration.

To evaluate the effectiveness of GroundingV2X, we conduct extensive experiments on three collaborative perception datasets, including one synthetic dataset, OPV2V (Xu et al., 2022c), and two real-world datasets, DAIR-V2X (Yu et al., 2022) and V2V4Real (Xu et al., 2023b). The results demonstrate that GroundingV2X achieves performance competitive with state-of-the-art methods in multi-model, single-task settings, while outperforming existing methods in multi-task scenarios. Meanwhile, its text-based communication protocol substantially reduces the communication overhead and the lightweight text encoder preserves high inference efficiency, making it a practical solution for real-world deployment.

2 Related Work

Collaborative Perception. Collaborative perception extends the sensing capability of individual agents (Liu et al., 2023). Existing approaches generally fall into early fusion, intermediate fusion and late fusion, among which intermediate fusion achieves a favorable balance among performance, bandwidth and robustness (Lin et al., 2024; Liu et al., 2020b;a; Hu et al., 2023; Liu et al., 2025). V2VNet (Wang et al., 2020) communicates compressed intermediate representations and aggregates them via graph neural networks, while V2X-ViT (Xu et al., 2022b) leverages a transformer-based module to model inter-agent interactions. Recent studies further improve communication efficiency by selectively transmitting spatially informative representations, such as Where2comm (Hu et al., 2022) and How2comm (Yang et al., 2023). Although achieving substantial improvements over single-agent perception, these approaches are primarily developed under homogeneous settings.

Heterogeneous Collaboration. Real-world collaborative systems inevitably involve heterogeneous agents equipped with different sensors, model architectures, perception tasks, etc., leading to substantial gaps during feature-level fusion (Wu et al., 2025; Chang et al., 2023). Recent studies have explored heterogeneous collaboration from different perspectives. HEAL (Lu et al., 2024) estab-

lishes a unified feature space and aligns newly introduced agents through backward alignment, while MPDA (Xu et al., 2022a), PnPDA (Luo et al., 2024), and PHCP (Si et al., 2025) introduce plug-and-play designs to bridge the heterogeneous feature spaces. STAMP (Gao et al., 2025b) further extends the heterogeneity toward task-agnostic settings by adapting and reverting the intermediate representations into and from a shared protocol feature space. More recently, GT-Space (Wang et al., 2026b) proposes a ground-truth derived feature space for better cross-agent alignment. Despite their effectiveness, these methods fundamentally restore compatibility among heterogeneous representations through collaborative post-training and feature alignment. In contrast, GroundingV2X establishes grounded text as a unified communication interface to avoid post-training or feature alignment.

Language-Guided Collaboration. Recent advances in large language models (LLMs) have motivated the use of natural language as a flexible medium for multi-agent collaboration (Wei et al., 2025). V2V-LLM (Chiu et al., 2025) formulates collaboration as a question-answering problem for grounding, object identification, and planning, while V2X-VLM (You et al., 2026) leverages vision-language reasoning for collaborative scene understanding and trajectory planning. LangCoop (Gao et al., 2025a) further adopts natural language as the inter-agent communication medium for high-level reasoning, planning, and control. However, existing language-based collaboration methods largely remain at the question-answering or high-level reasoning stage, and heavily rely on large-scale language models. In contrast, GroundingV2X encodes the shared grounded text with a lightweight text encoder and explicitly integrates the text features into the perceptual representation.

Visual Grounding. Visual grounding aims to establish fine-grained correspondences between linguistic concepts and visual regions (Xiao et al., 2025; Wang et al., 2026a). MDETR (Kamath et al., 2021) performs text-conditioned object detection through joint vision-language reasoning, while GLIP (Li et al., 2022) unifies object detection and phrase grounding via grounded language-image pre-training. Grounding DINO (Liu et al., 2024) further strengthens vision-language grounding for open-set object detection. Different from these methods that localize objects matching linguistic queries within observed visual context, GroundingV2X injects additional grounded information of those unobserved objects or regions into the local perceptual representation.

3 METHOD

3.1 PROBLEM FORMULATION AND FRAMEWORK OVERVIEW

We consider a collaborative perception scenario with N heterogeneous agents $\mathcal{A} = \{\mathrm{A}_i\}_{i=1}^{N}$, where each agent is equipped with distinct sensor equipment, model architectures and perception tasks. For an agent A_i, it first extracts the intermediate feature $F_i \in \mathbb{R}^{H_i \times W_i \times C_i}$ from its sensor input X_i and then predicts its perceptual results Y_i through its task-specific decoder D_i, where H_i, W_i, and C_i denote the height, width, and channel dimensions of the intermediate feature.

Instead of sharing and aggregating the intermediate feature, GroundingV2X introduces grounded text as a unified communication interface for heterogeneous collaboration, which is illustrated in Figure 2. In this framework, each agent converts its perceptual results into grounded text and only exchanges the grounded text with other agents. Agents receiving grounded text from others extract the text features using a lightweight text encoder with Fourier numerical embedding and sub-sentence level attention, and selectively integrate the text features into their spatially relevant intermediate representation through a position-aware text-to-feature enhancer. In addition, each agent is pre-trained independently using its perceived grounded text, synthetic grounded text, and annotation-derived pseudo-labels to improve its generalization to diverse textual descriptions.

3.2 GROUNDED TEXT CONSTRUCTION AND COMMUNICATION

According to the specific perception tasks such as 3D object detection and Bird's Eye View (BEV) segmentation, the perceptual results of each agent can be converted into object-level or region-level descriptions. Each detected object or region is described by one individual sentence, while the entire grounded text consists of multiple sentences, which can be formulated as

$$\mathbb{T}_{j\to i} = (S^r_{j\to i})_{r=1}^{R_{j\to i}}, \qquad S^r_{j\to i} = \big(c^r, \kappa^r, (k^r_m, v^r_m)_{m=1}^{m_r}\big), \tag{1}$$

where $\mathbb{T}_{j\to i}$ denotes the grounded text transmitted from agent A_j to agent A_i, $S^r_{j\to i}$ is the r-th sentence describing an object or region, and $R_{j\to i}$ represents the number of sentences. In each sen-

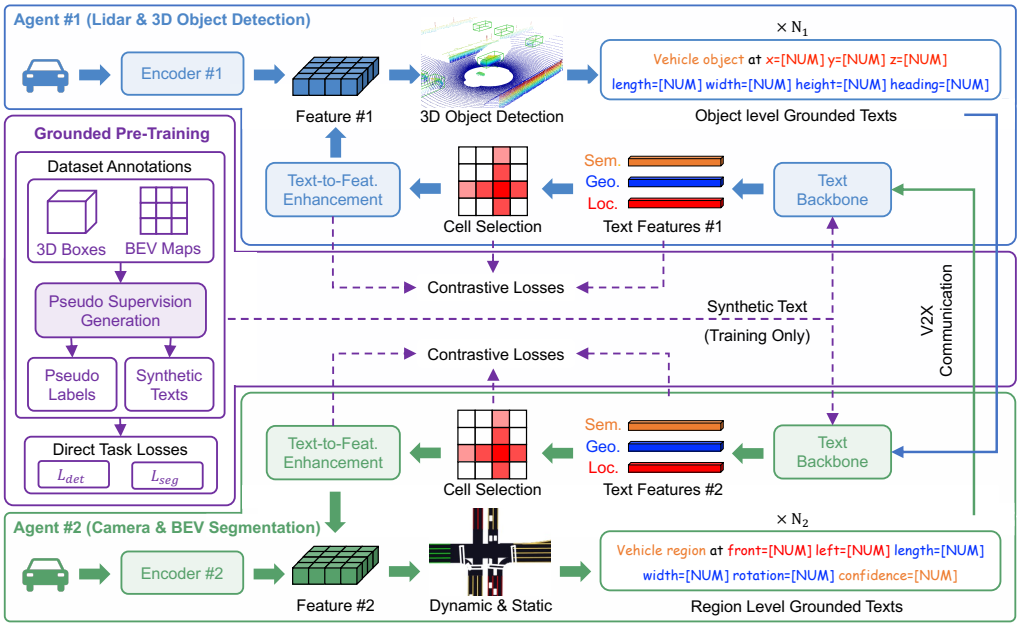


Figure 2: **Overview of GroundingV2X.** Agents with distinct sensors, models and tasks extract intermediate features, perform perception and construct the corresponding grounded text. Agents receiving grounded text encode it using Fourier numerical embeddings with the `[NUM]` tokens and specialized sub-sentence level attention. The encoded text features are then selectively integrated into the spatially relevant perceptual features through a position-aware text-to-feature enhancer. Grounded pre-training constructs diverse synthetic grounded texts and pseudo-labels from dataset annotations for generalized text understanding.

tence, c^r specifies the semantic category, κ^r distinguishes objects from regions, and m_r counts the key-value fields. The keys identify positional information (e.g., x/y or front/left), geometry (length, width, height), and semantic attributes such as confidence, while the values preserve the corresponding numerical details. Specifically, the objects or regions are described from the receiver's perspective during text construction. For instance, an abbreviated object description can be "a vehicle object, front=52.5, right=2.2, length=4.7, confidence=0.88", while the region description can be constructed in the same manner with different keys. Such language-based descriptions are used as a unified interface for communication among heterogeneous agents without requiring compatible intermediate representations.

3.3 Fourier Numerical and Sub-Sentence Level Text Encoding

To preserve numerical values without fragmenting them into ordinary subword tokens, we replace each number v with a special `[NUM]` token and encode its value using Fourier number embedding (Zhou et al., 2026):

$$\phi(v;T) = \left[\cos\left(\frac{2\pi v}{T}\right), \sin\left(\frac{2\pi v}{T}\right)\right], \qquad \mathrm{FoNE}(v) = \bigoplus_{p=1}^{P} \phi(v;T_p), \tag{2}$$

where $\bigoplus$ denotes concatenation and T_p is the period coefficient that determines the frequency of the Fourier basis. Rather than using decimal periods as in the original FoNE, we use irrational periods $T_p = T_{\min}\rho^{p-1}$ with $\rho = (1+\sqrt{5})/2$, spanning the required numerical resolution and increasing the precision of the differentiation between similar numerical values. The tokenization of number v integrates the Fourier number embedding with the `[NUM]` token embedding:

$$e_i(v) = e_i^{[\mathrm{NUM}]} + \mathrm{Linear}\left(\mathrm{FoNE}(v)\right). \tag{3}$$

Then the tokenized grounded text is processed by a lightweight text encoder with a sub-sentence level attention mask, where tokens can only attend to other tokens within the same sentence while

cross-sentence attention is masked. Context therefore associates each numerical value with its key without mixing unrelated entities.

3.4 Position-Aware Text-to-Feature Enhancement

For the receiving agent, we index all incoming sentences by r and denote their encoded text features by $Z_r \in \mathbb{R}^{L_r \times d}$, which are then projected into positional $Z_r^{pos} \in \mathbb{R}^{L_r \times C}$ and contextual $Z_r^{ctx} \in \mathbb{R}^{L_r \times C}$ spaces to locate the text features and capture the semantic-geometric context, respectively.

To be specific, for the receiver agent A_i, a positional query Q_i is constructed through anchor parameters for detection or center coordinates for BEV segmentation. Then the token-cell cosine similarities are computed between the positional query Q_i and each token of the positional text features Z_r^{pos}, and averaged within each sentence:

$$S^r(h,w) = \frac{1}{L_r} \sum_{\ell=1}^{L_r} \cos\Big(Q_{i,h,w}, Z_{r,\ell}^{pos}\Big), \qquad \cos(a,b) = \frac{a^\top b}{\|a\|_2 \|b\|_2}, \tag{4}$$

where $S^r \in \mathbb{R}^{H_i \times W_i}$ represents the averaged similarity score between the tokens of sentence S^r and a specific cell (h,w) of the positional query and L_r denotes the token number of sentence S^r.

For a sentence S^r, its K highest-scoring cells are selected to form the eligible set $\mathcal{B}_r$. At each cell, eligible sentences compete with a learned invalid option:

$$w_r(h,w) = \frac{\mathbb{I}[(h,w) \in \mathcal{B}_r] \exp(S^r(h,w)/\tau)}{\exp(b_0/\tau) + \sum_t \mathbb{I}[(h,w) \in \mathcal{B}_t] \exp(S^t(h,w)/\tau)}, \tag{5}$$

where $\tau > 0$ is a learned temperature, b_0 denotes the learned invalid score and $\sum_t$ denotes the sum over all sentences t. The invalid option suppresses weak matches, while only the cells and sentences with $w_r(h,w) > \epsilon$ are retained for further feature enhancement, where ϵ is a small threshold to filter out low-confidence matches.

The retained positional query then attends to the contextual tokens of the selected sentences through cross-attention with a query conditioned on the cell's relative offset in the positional query space. Let $\bar{Z}_r^{pos} = \frac{1}{L_r} \sum_{\ell=1}^{L_r} Z_{r,\ell}^{pos}$ denote the token-averaged positional vector of sentence S^r. Then

$$\begin{aligned} \xi_{r,h,w} &= \mathbb{E} + \text{Linear}\left(Q_{i,h,w} - \bar{Z}_r^{pos}\right), \\ u_{r,h,w} &= \text{CrossAttn}\big(\xi_{r,h,w}, Z_r^{ctx}, Z_r^{ctx}\big), \end{aligned} \tag{6}$$

where $\mathbb{E}$ is a learned embedding and $u_{r,h,w}$ denotes the context of sentence S^r at cell (h,w). The resulting contextual features are weighted by $w_r(h,w)$ and scattered back to the corresponding cells:

$$\widetilde{F}_i = F_i + \sum_r \text{Scatter}_{\Omega_r}\big([w_r(h,w) u_{r,h,w}]_{(h,w)\in\Omega_r}\big), \tag{7}$$

where $\Omega_r = \{(h,w) : w_r(h,w) > \epsilon\}$ is the set of selected cells for sentence S^r, and the scattered feature is then refined through a residual module with linear projection and convolution to obtain the final enhanced features for perception decoding.

3.5 Grounded Pre-Training and Learning Objectives

Each agent is pre-trained independently with its perception encoder frozen. However, heterogeneous agents may produce diverse textual expressions. To enable the model to generalize across different agents without post-training or task-specific adaptation, we introduce a grounded pre-training strategy that leverages both perceived and synthetic grounded texts, as well as annotation-derived pseudo-labels for supervision.

Specifically, each agent applies its own model to compatible collaborator-view inputs to generate perceived grounded text from multiple perspectives, resulting in self-collaborative training. In addition, we synthesize auxiliary grounded text and pseudo-label pairs across different perception tasks in unobserved regions based on the dataset annotations, enabling the model to learn diverse object or region descriptions in distinct positions. Moreover, during the pre-training, we simulate the

imperfections by applying synonymous keywords and permuted key-value order to the grounded text, as well as randomly dropping some sentences, which exposes the model to potential errors in real-world communication and enhances its robustness.

During grounded pre-training, each agent A_i is optimized individually with native task supervision, positional grounding, numerical binding, and task-specific auxiliary decoding.

Native task supervision. The native task loss $\mathcal{L}_{task,i}$ supervises predictions decoded from the enhanced features $\widetilde{F}_i$ with each agent's own detection or segmentation losses.

Positional grounding. Each token of sentence S^r in the positional space Z_r^{pos} is aligned with the footprint $\mathcal{P}_r$, i.e., the cells of the positional query Q_i covered by the objects or regions stated in the grounded text. Since one positional key-value pair constrains only one axis, x-coordinate tokens $\mathcal{I}_{r,x}$ take all cells of the BEV grid lying in the columns spanned by $\mathcal{P}_r$ as positives $\mathcal{P}_{r,x}$, y-coordinate tokens $\mathcal{I}_{r,y}$ take all cells in the rows spanned by $\mathcal{P}_r$ as $\mathcal{P}_{r,y}$, and the remaining tokens $\mathcal{I}_{r,o}$ take exactly the footprint, $\mathcal{P}_{r,o} = \mathcal{P}_r$.

With $\mathcal{P}_{r,\ell} \in \{\mathcal{P}_{r,x}, \mathcal{P}_{r,y}, \mathcal{P}_{r,o}\}$ as the positive set of token ℓ according to the group $\mathcal{I}_{r,x}$, $\mathcal{I}_{r,y}$, or $\mathcal{I}_{r,o}$ it belongs to, and $p_{r,\ell}(h,w)$ as the softmax over all BEV cells of the token-to-cell similarities in Eq. (4), we apply a multi-positive InfoNCE loss (van den Oord et al., 2018) that aligns each key-value token toward its positive cells $\mathcal{P}_{r,\ell}$ and against all other cells:

$$\mathcal{L}_{pos,i} = -\frac{1}{\sum_r L_r} \sum_r \sum_{\ell=1}^{L_r} \frac{1}{|\mathcal{P}_{r,\ell}|} \sum_{(h,w)\in\mathcal{P}_{r,\ell}} \log p_{r,\ell}(h,w), \tag{8}$$

which supervises the spatial support of each key-value pair rather than the entire sentence.

Numerical binding. Each `[NUM]` token in the contextual space Z_r^{ctx} is aligned with the value it carries to preserve the numerical details. Each value v is mapped to an embedding $\psi_i(v) = \mathrm{Linear}(\mathrm{FoNE}(v))$, and an InfoNCE loss aligns the numerical tokens in Z_r^{ctx} toward the embeddings of their values and away from the embeddings of the other values:

$$\mathcal{L}_{val,i} = -\frac{1}{L_v} \sum_{n=1}^{L_v} \log \frac{\exp(\cos(Z_n^{ctx}, \psi_i(v_n))/\tau_v)}{\sum_{m\in\mathcal{J}_n} \exp(\cos(Z_n^{ctx}, \psi_i(v_m))/\tau_v)}, \tag{9}$$

where L_v is the number of `[NUM]` tokens, $\tau_v > 0$ a temperature, and $\mathcal{J}_n = \{n\} \cup \{m : |v_m - v_n| \geq \delta\}$ restricts the negatives to tokens whose values differ from v_n by at least δ, excluding near-duplicate values with nearly identical embeddings.

Task-specific auxiliary decoding. For agents whose task decoder D_i predicts multiple attributes stated in one sentence, e.g., box geometry, confidence, and heading in detection, an auxiliary decoding loss $\mathcal{L}_{aux,i}$ is imposed by applying D_i to the composed context $u_{r,h,w}$ in Eq. (6), thereby enforcing that the context encodes the stated attributes. For instance, an agent performing object detection applies its classification, regression, and heading decoders to the composed context $u_{r,h,w}$, which is further supervised by the box attributes stated in the sentence S^r.

Overall objective. The overall loss of agent A_i is the weighted sum

$$\mathcal{L}_i = \mathcal{L}_{task,i} + \lambda_{pos}\mathcal{L}_{pos,i} + \lambda_{val}\mathcal{L}_{val,i} + \lambda_{aux}\mathcal{L}_{aux,i}, \tag{10}$$

where λ_{pos}, λ_{val}, and λ_{aux} are the weighting coefficients of the corresponding terms.

4 Experiments

4.1 Experimental Settings

Datasets and metrics. We evaluate GroundingV2X on one simulated dataset (OPV2V) and two real-world datasets (V2V4Real and DAIR-V2X), covering various scenarios and conditions like vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) collaboration. For perception performance, we adopt average precision (AP) as the evaluation metric for 3D object detection and mean intersection-over-union (mIoU) for BEV segmentation. In addition to perception performance, we

Table 1: **Performance comparison of model-heterogeneous collaboration on OPV2V.** With $\mathrm{A}_{l,p}^{det}$ as the ego agent, 3 heterogeneous collaborators are cumulatively added. AP@0.5 and AP@0.7 are reported for accuracy comparison, while Avg. ΔComm. (MB/agent) and Avg. ΔTFLOPs (TFLOPs/agent) report the incremental communication and computational overhead relative to no collaboration, normalized by the number of collaborators.

Methods	AP@0.5↑			AP@0.7↑			Avg. ΔComm. ↓	Avg. ΔTFLOPs ↓
Ego Agent $\mathrm{A}_{l,p}^{det}$	$+\,\mathrm{A}_{c,r}^{det}$	$+\,\mathrm{A}_{l,s}^{det}$	$+\,\mathrm{A}_{c,e}^{det}$	$+\,\mathrm{A}_{c,r}^{det}$	$+\,\mathrm{A}_{l,s}^{det}$	$+\,\mathrm{A}_{c,e}^{det}$	MB/Agent	T/Agent
No Collaboration	0.748	0.748	0.748	0.606	0.606	0.606	0	0
Late fusion	0.775	0.833	0.834	0.599	0.685	0.685	$\simeq 0$	0.105
V2XViT	0.822	0.888	0.882	0.655	0.765	0.753	128	0.353
CoBEVT	0.822	0.885	0.885	0.671	0.742	0.742	128	0.281
HM-ViT	0.813	0.871	0.876	0.646	0.743	0.755	128	0.336
HEAL	0.826	0.892	0.894	0.726	0.812	0.813	128	0.105
STAMP	0.796	0.852	0.861	0.695	0.779	0.791	128	0.127
GroundingV2X	0.842	0.901	0.901	0.743	0.813	0.813	$\simeq 0$	0.183

Table 2: **Performance comparison of model-heterogeneous collaboration on V2V4Real and DAIR-V2X.** V2V4Real pairs two vehicles (V2V), while DAIR-V2X pairs a vehicle as ego agent and an infrastructure as collaborator (V2I). The first agent listed is the ego. We report AP@0.3, AP@0.5, ΔComm. and ΔTFLOPs for accuracy and efficiency comparison.

Datasets	V2V4Real			DAIR-V2X		
Agent Settings	$\mathrm{A}_{l,p}^{det}$ + $\mathrm{A}_{l,s}^{det}$			$\mathrm{A}_{l,s}^{det}$+ $\mathrm{A}_{l,p}^{det}$		
Metric	AP@0.3 ↑	ΔComm.↓	ΔTFLOPs↓	AP@0.5↑	ΔComm.↓	ΔTFLOPs↓
No Collaboration	0.647	0	0	0.587	0	0
Late fusion	0.741	$\simeq 0$	0.038	0.594	$\simeq 0$	0.149
HEAL	0.811	32	0.038	0.770	128	0.149
STAMP	0.753	32	0.028	0.758	128	0.125
GroundingV2X	0.790	$\simeq 0$	0.055	0.655	$\simeq 0$	0.283

evaluate communication efficiency and computational cost using communication volume (Comm.) and floating-point operations (FLOPs), respectively.

Implementation details. We specify agents with different sensor modalities (l: LiDAR; c: camera), encoders, and tasks (det: object detection; bev: segmentation). Encoders include PointPillars (p) (Lang et al., 2019) and SECOND (s) (Yan et al., 2018) for LiDAR input, and ResNet (r) (He et al., 2016) and EfficientNet (e) (Tan & Le, 2019) for camera input. For instance, an agent equipped with a LiDAR sensor and a SECOND encoder performing 3D object detection is denoted as $\mathrm{A}_{l,s}^{det}$. Each single-agent model is trained independently for 30 epochs with Adam at an initial learning rate of 10^{-3}. We then freeze its perception encoder and train the text encoder, text-to-feature enhancer and task decoders for 30 epochs with a learning rate of 10^{-4} and a gradient clipping of 0.1. We adopt BERT (Devlin et al., 2019) as the pre-trained text backbone whose learning rate is scaled by 0.1. Independently trained agents are jointly deployed for evaluation.

4.2 Model-Agnostic Heterogeneous Collaboration

We first consider model-heterogeneous collaboration scenarios where all agents perform 3D object detection but adopt different model architectures and sensor inputs.

Table 1 presents the performance comparison on OPV2V with several representative and state-of-the-art methods. Following the evaluation protocol in HEAL (Lu et al., 2024), one agent is selected as the ego agent and additional agents are progressively introduced. As we can see, GroundingV2X consistently achieves a superior performance compared to state-of-the-art methods. In addition,

Table 3: **Performance comparison of model- and task-heterogeneous collaboration on OPV2V.** Four agents differ in sensors, encoders, sensing ranges, and tasks. Each serves as the ego in turn with the other three as collaborators. We report AP@0.5 for object detection and mIoU for BEV segmentation. Homogeneity applies the ego agent's model and task for all agents.

Ego Agent	$A^{det}_{l,p}$	$A^{det}_{c,r}$	$A^{bev}_{l,s}$	$A^{bev}_{c,e}$
Semantic Category	Vehicle	Vehicle	Dynamic	Static
Sensing Range (m)	$[\pm 204.8, \pm 102.4]$	$[\pm 51.2, \pm 51.2]$	$[\pm 51.2, \pm 51.2]$	$[\pm 51.2, \pm 51.2]$
Metric	AP@0.5↑	AP@0.5↑	mIoU↑	mIoU↑
No Collaboration	0.748	0.338	0.519	0.442
Homogeneity	0.957	0.566	0.639	0.479
STAMP	0.736	0.482	0.152	0.153
GroundingV2X	0.849	0.705	0.609	0.477

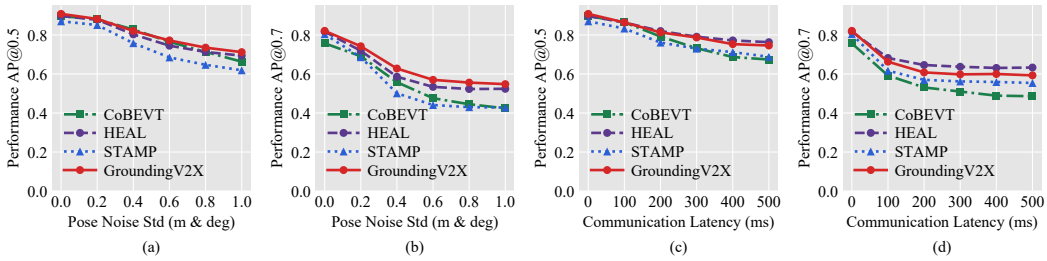


Figure 3: **Robustness to pose noise and communication latency on OPV2V.** (a) and (b) report AP@0.5 and AP@0.7 under different levels of Gaussian pose noise $\mathcal{N}(0, \sigma^2)$ with σ from 0 to 1.0 m. (c) and (d) present AP@0.5 and AP@0.7 under communication latency from 0 to 500 ms.

GroundingV2X significantly reduces the communication overhead compared to intermediate fusion methods, while maintaining a moderate computational cost of 0.183 TFLOPs for each collaborator.

Table 2 presents the results of two-agent real-world collaboration on the V2V4Real and DAIR-V2X datasets. On V2V4Real, GroundingV2X consistently achieves a competitive performance of AP@0.3 with state-of-the-art methods. However, GroundingV2X underperforms feature-level fusion in DAIR-V2X at AP@0.5, due to the annotation mismatch between two agents, which can be inferred from the limited improvement in late fusion while feature-level fusion can partially absorb such mismatches through the feature alignment and joint supervision.

4.3 Model- and Task-Agnostic Heterogeneous Collaboration

We then evaluate joint model and task heterogeneity with two detection agents and two BEV segmentation agents. Their sensing ranges, sensor inputs and encoders are also different, testing whether grounded text can bridge such challenging heterogeneous settings.

Table 3 presents the results of model- and task-heterogeneous collaboration on OPV2V. We compare the performance of GroundingV2X with the state-of-the-art method STAMP and the homogeneity setting, where all agents share the same model and task. As we can see, compared to STAMP, GroundingV2X achieves consistently better performance and effective collaboration gains across four agents, demonstrating its effectiveness in handling both model and task heterogeneity. The results highlight the superiority of grounded text as a unified interface.

4.4 Robustness to Pose Noise and Communication Latency

In practical V2X communication scenarios, the pose estimates of different agents may be inaccurate, and communication latency is inevitable. Therefore, we further evaluate the robustness of GroundingV2X to pose noise and communication latency.

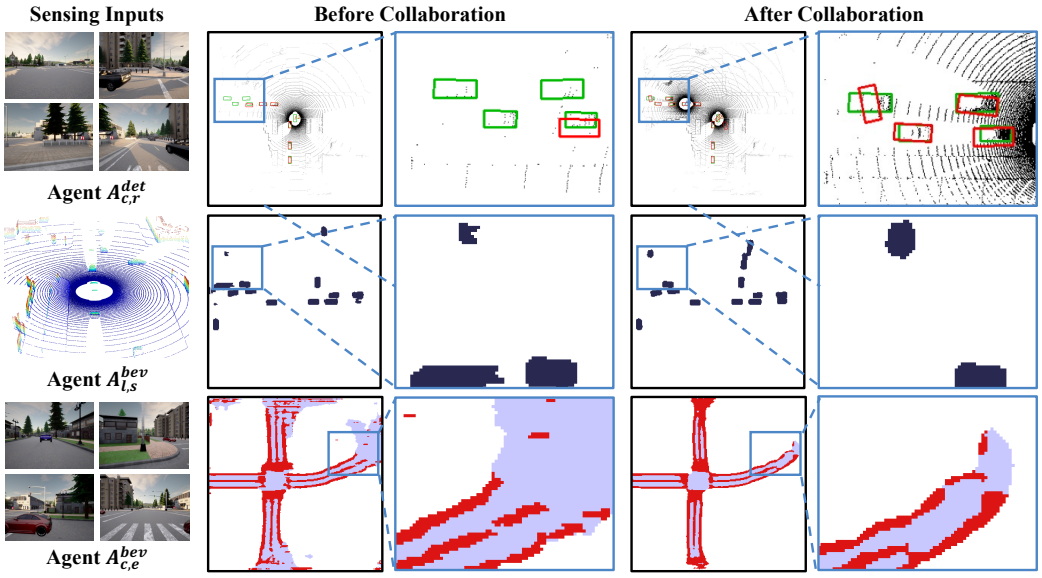


Figure 4: **Qualitative visualization** of the perception results of 3 heterogeneous agents with different sensor modalities and tasks. We illustrate their sensor inputs along with their perception outputs before and after collaboration using GroundingV2X. The enlarged regions highlight the recovered detections and more accurate BEV segmentation.

Pose noise. We add Gaussian pose noise $\mathcal{N}(0, \sigma^2)$ with standard deviation varying from 0.0 to 1.0 m to the agent's pose estimation. The results of AP@0.5 and AP@0.7 under different levels of pose noise are presented in Figure 3 (a) and (b). As we can see, GroundingV2X demonstrates superior robustness to pose noise compared to other methods, maintaining higher accuracy with increasing levels of pose noise.

Communication latency. We simulate communication latency by introducing delays in evaluation samples of the collaborators and compare the performance of GroundingV2X with other methods under different latency levels ranging from 0 to 500 milliseconds. Figure 3 (c) and (d) present the results of AP@0.5 and AP@0.7 with increasing communication latency. As we can see, GroundingV2X maintains relatively stable performance as communication latency increases, slightly below HEAL's. This gap suggests that grounded text could readily carry temporal attributes such as timestamps to compensate for the latency, though they are not included in the current implementation.

4.5 Qualitative Visualization

Figure 4 illustrates the sensor inputs along with the perception outputs before and after collaboration of three heterogeneous agents that perform object detection, dynamic BEV segmentation and static BEV segmentation, respectively. The enlarged regions highlight the recovered object detections and more complete BEV segmentation. These qualitative visualizations demonstrate that each agent can effectively gain complementary information from other agents through grounded text.

5 Conclusion

We presented GroundingV2X, enabling heterogeneous agents to collaborate through grounded text without cross-agent feature alignment or collaborative post-training. Numerically aware text encoding, sub-sentence level attention and position-aware feature enhancement efficiently integrate the received information. Grounded pre-training with annotation-derived synthetic text and pseudo-labels allows agents to learn this interface independently. Experiments demonstrate competitive performance with substantially lower communication volume and low computational overhead, establishing grounded text as an effective interface for heterogeneous collaborative perception.

AI USE STATEMENT

We have used AI tools to assist in drafting portions of the manuscript, language polishing and implementing code for specific functions. We have carefully reviewed and revised the AI-assisted outputs and take full responsibility for the content in this work.

REPRODUCIBILITY STATEMENT.

To ensure the reproducibility of our results, we have provided detailed implementation information, including the grounded text grammar, training configurations, and hyperparameters in Appendix A. We will also release the code and pre-trained weights to facilitate further research in this area.

REFERENCES

Zhengwei Bai, Guoyuan Wu, Matthew J Barth, Yongkang Liu, Emrah Akin Sisbot, Kentaro Oguchi, and Zhitong Huang. A survey and framework of cooperative perception: From heterogeneous singleton to hierarchical cooperation. *IEEE Transactions on Intelligent Transportation Systems*, 25(11):15191–15209, 2024.

Cheng Chang, Jiawei Zhang, Kunpeng Zhang, Wenqin Zhong, Xinyu Peng, Shen Li, and Li Li. Bev-v2x: Cooperative birds-eye-view fusion and grid occupancy prediction via v2x-based data sharing. *IEEE Transactions on Intelligent Vehicles*, 8(11):4498–4514, 2023.

Hsu-kuang Chiu, Ryo Hachiuma, Chien-Yi Wang, Stephen F Smith, Yu-Chiang Frank Wang, and Min-Hung Chen. V2V-LLM: Vehicle-to-vehicle cooperative autonomous driving with multi-modal large language models. *arXiv preprint arXiv:2502.09980*, 2025.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Xiangbo Gao, Yuheng Wu, Rujia Wang, Chenxi Liu, Yang Zhou, and Zhengzhong Tu. LangCoop: Collaborative driving with language. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4226–4237. IEEE, 2025a.

Xiangbo Gao, Runsheng Xu, Jiachen Li, Ziran Wang, Zhiwen Fan, and Zhengzhong Tu. STAMP: Scalable task- and model-agnostic collaborative perception. In *International Conference on Learning Representations*, 2025b.

Siavash Golkar, Mariel Pettee, Michael Eickenberg, Alberto Bietti, Miles Cranmer, Geraud Krawezik, Francois Lanusse, Michael McCabe, Ruben Ohana, Liam Holden Parker, et al. xval: A continuous number encoding for large language models. In *NeurIPS 2023 AI for Science Workshop*, 2023.

Yushan Han, Hui Zhang, Huifang Li, Yi Jin, Congyan Lang, and Yidong Li. Collaborative perception in autonomous driving: Methods, datasets, and challenges. *IEEE Intelligent Transportation Systems Magazine*, 15(6):131–151, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Yue Hu, Shaoheng Fang, Zixing Lei, Yiqi Zhong, and Siheng Chen. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. *Advances in neural information processing systems*, 35:4874–4886, 2022.

Yue Hu, Yifan Lu, Runsheng Xu, Weidi Xie, Siheng Chen, and Yanfeng Wang. Collaboration helps camera overtake lidar in 3D detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9243–9252, 2023.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. MDETR-modulated detection for end-to-end multi-modal understanding. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 1760–1770. IEEE, 2021.

Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. PointPillars: Fast encoders for object detection from point clouds. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 12689–12697. IEEE, 2019.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10955–10965. IEEE, 2022.

Chunmian Lin, Daxin Tian, Xuting Duan, Jianshan Zhou, Dezong Zhao, and Dongpu Cao. V2VFormer: Vehicle-to-vehicle cooperative perception with spatial-channel transformer. *IEEE Transactions on Intelligent Vehicles*, 9(2):3384–3395, 2024. doi: 10.1109/TIV.2024.3353254.

Haizhuang Liu, Huazhen Chu, Junbao Zhuo, Bochao Zou, Jiansheng Chen, and Huimin Ma. SparseComm: An efficient sparse communication framework for vehicle-infrastructure cooperative 3D detection. *Pattern Recognition*, 158:110961, 2025.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

Si Liu, Chen Gao, Yuan Chen, Xingyu Peng, Xianghao Kong, Kun Wang, Runsheng Xu, Wentao Jiang, Hao Xiang, Jiaqi Ma, et al. Towards vehicle-to-everything autonomous driving: A survey on collaborative perception. *arXiv preprint arXiv:2308.16714*, 2023.

Yen-Cheng Liu, Junjiao Tian, Nathaniel Glaser, and Zsolt Kira. When2com: Multi-agent perception via communication graph grouping. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 4106–4115, 2020a.

Yen-Cheng Liu, Junjiao Tian, Chih-Yao Ma, Nathan Glaser, Chia-Wen Kuo, and Zsolt Kira. Who2com: Collaborative perception via learnable handshake communication. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6876–6883. IEEE, 2020b.

Yifan Lu, Yue Hu, Yiqi Zhong, Dequan Wang, Yanfeng Wang, and Siheng Chen. An extensible framework for open heterogeneous collaborative perception. In *International Conference on Learning Representations*, 2024.

Tianyou Luo, Quan Yuan, Guiyang Luo, Yuchen Xia, Yujia Yang, and Jinglin Li. Plug and play: A representation enhanced domain adapter for collaborative perception. In *European Conference on Computer Vision*, pp. 287–303. Springer, 2024.

Hao Si, Ehsan Javanmardi, and Manabu Tsukada. You share beliefs, i adapt: Progressive heterogeneous collaborative perception. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 27521–27530. IEEE, 2025.

Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PmLR, 2019.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Jiacong Wang, Zijian Kang, Haochen Wang, Xiao Liang, Ya Wang, Jiawen Li, Bohong Wu, Jiao Ran, Haiyong Jiang, Chao Feng, et al. Vgr: Visual grounded reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 89506–89528, 2026a.

Tsun-Hsuan Wang, Sivabalan Manivasagam, Ming Liang, Bin Yang, Wenyuan Zeng, and Raquel Urtasun. V2VNet: Vehicle-to-vehicle communication for joint perception and prediction. In *European conference on computer vision*, pp. 605–621. Springer, 2020.

Wentao Wang, Haoran Xu, and Guang Tan. GT-Space: Enhancing heterogeneous collaborative perception with ground truth feature space. In *International Conference on Learning Representations*, 2026b.

Chuheng Wei, Guoyuan Wu, and Matthew J Barth. Cooperative perception for automated driving: A survey of algorithms, applications, and future directions. *Proceedings of the IEEE*, 2025.

Hanlin Wu, Pengfei Lin, Ehsan Javanmardi, Naren Bao, Bo Qian, Hao Si, and Manabu Tsukada. A synthetic benchmark for collaborative 3d semantic occupancy prediction in v2x autonomous driving. *arXiv preprint arXiv:2506.17004*, 2, 2025.

Linhui Xiao, Xiaoshan Yang, Xiangyuan Lan, Yaowei Wang, and Changsheng Xu. Towards visual grounding: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Runsheng Xu, Jinlong Li, Xiaoyu Dong, Hongkai Yu, and Jiaqi Ma. Bridging the domain gap for multi-agent perception. *arXiv preprint arXiv:2210.08451*, 2022a.

Runsheng Xu, Hao Xiang, Zhengzhong Tu, Xin Xia, Ming-Hsuan Yang, and Jiaqi Ma. V2X-ViT: Vehicle-to-everything cooperative perception with vision transformer. In *European conference on computer vision*, pp. 107–124. Springer, 2022b.

Runsheng Xu, Hao Xiang, Xin Xia, Xu Han, Jinlong Li, and Jiaqi Ma. OPV2V: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 2583–2589, 2022c. doi: 10.1109/ICRA46639.2022.9812038.

Runsheng Xu, Zhengzhong Tu, Hao Xiang, Wei Shao, Bolei Zhou, and Jiaqi Ma. CoBEVT: Cooperative bird's eye view semantic segmentation with sparse transformers. In Karen Liu, Dana Kulic, and Jeff Ichnowski (eds.), *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 989–1000. PMLR, 14–18 Dec 2023a.

Runsheng Xu, Xin Xia, Jinlong Li, Hanzhao Li, Shuo Zhang, Zhengzhong Tu, Zonglin Meng, Hao Xiang, Xiaoyu Dong, Rui Song, et al. V2V4Real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13712–13722, 2023b.

Yan Yan, Yuxing Mao, and Bo Li. SECOND: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018.

Dingkang Yang, Kun Yang, Yuzheng Wang, Jing Liu, Zhi Xu, Rongbin Yin, Peng Zhai, and Lihua Zhang. How2comm: Communication-efficient and collaboration-pragmatic multi-agent perception. *Advances in Neural Information Processing Systems*, 36:25151–25164, 2023.

Junwei You, Zhuoyu Jiang, Zilin Huang, Haotian Shi, Rui Gan, Keshu Wu, Xi Cheng, Xiaopeng Li, and Bin Ran. V2X-VLM: End-to-end V2X cooperative autonomous driving through large vision-language models. *Transportation Research Part C: Emerging Technologies*, 183:105457, 2026.

Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui Yuan, et al. DAIR-V2X: A large-scale dataset for vehicle-infrastructure cooperative 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21361–21370, 2022.

Tianyi Zhou, Deqing Fu, Mahdi Soltanolkotabi, Robin Jia, and Vatsal Sharan. FoNE: Precise single-token number embeddings via Fourier features. In *International Conference on Learning Representations*, 2026.

A Implementation Details

Architecture and gradient flow. The lightweight text encoder is BERT-base-uncased with 768-dimensional tokens, followed by 256-dimensional projections into the positional and contextual spaces. The positional query Q_i is obtained by encoding the anchor parameters, or the cell centers for BEV segmentation, with the Fourier number embedding and a two-layer MLP with 128 hidden units. The cross-attention in the text-to-feature enhancer uses four heads, and the residual module refining the enhanced features $\widetilde{F}_i$ consists of a linear projection followed by two 3×3 convolutions with GroupNorm and GELU. The perception encoder is frozen during grounded pre-training. The similarity scores S^r are detached from $\mathcal{L}_{task,i}$ so that the positional projection is trained only by $\mathcal{L}_{pos,i}$, and $\mathcal{L}_{aux,i}$ detaches the parameters of the task decoder D_i while keeping gradients to the composed context $u_{r,h,w}$.

Table 4: Key-value fields of the sentences for object and region descriptions.

Sentence Type	Keys
`<category> object`	confidence, x, y, z, height, width, length, heading
`<category> region`	coverage, front/back, left/right, length, width, rotation

Grounded text format. Each sentence S^r starts with its semantic category c^r and structure κ^r, i.e., object or region description, followed by comma-separated key-value fields as summarized in Table 4. Positional and geometric values are in meters, angles in radians, and confidence and coverage in $[0, 1]$. Positional keys are written either as signed `x`/`y` or as directional words with nonnegative values, i.e., `front`/`back` for $\pm x$ and `right`/`left` for $\pm y$. The heading of an object keeps its facing direction, whereas the rotation of a region is represented modulo π since a region has no orientation. Before text encoding, numerical values are rounded to two decimal places and replaced by `[NUM]` tokens, and the values are kept for Fourier number embedding. Then the sentences of all objects and regions are joined via the `[SEP]` token.

Region description construction. Different from object sentences, which can be derived directly from the 3D bounding box results, region sentences are constructed from BEV segmentation results, which are composed of multiple discrete cells. For each semantic category, we use minimum-area oriented rectangles to enclose the connected cells of this semantic category into a region, with elongated regions split along their long axis, and we adopt coverage to represent the fraction of cells inside a region that are occupied by the same semantic category.

Supervision footprint construction. To align the sentences with the actual position, we need to establish a correspondence between the positional tokens of the sentences and the actual positions in the feature space. The footprint is determined by the task of the receiving agent regardless of the sentence type. For an agent with an anchor-based detector, the footprint $\mathcal{P}_r$ contains the cells whose anchors are assigned to the stated geometry under the detector's own label-assignment rule. For an agent performing BEV segmentation, it contains the cells whose centers fall inside the stated geometry.

Synthetic grounded text and pseudo-labels. Besides the perceived grounded text produced by the agent's own model, each training frame is supplemented with three kinds of synthetic sentences constructed from the dataset annotations:

- *Cross-task sentences of annotated objects/regions.* For the annotated objects or regions in the scene, we construct noisy sentences of the other task for generalized training. For instance, an agent performing object detection receives region sentences, while an agent performing BEV segmentation receives object sentences. These sentences are derived from the annotated objects or regions in the dataset. Meanwhile, part of the perceived sentences of the agent's own task that describe the same objects are randomly dropped, so that the cross-task sentences have to carry these objects or regions on their own. These sentences are then fused into $\widetilde{F}_i$ and supervised by the original task labels.
- *Sentences of virtual entities.* We further fabricate a few virtual detection or region targets per frame placed at positions where no object or region annotation exists. Their sentences are authored in the same way as annotated ones and take the form of either task, i.e., object or region sentences,

according to a fixed ratio, and are treated as part of the received grounded text for feature enhancement with the labels amended. Therefore, the agent is trained to ground text at arbitrary positions, rather than only at positions covered by the dataset annotations.

- *Alignment-only sentences.* To learn the alignment of grounded text at arbitrary positions, we sample object and region sentences uniformly over the sensing range with random positions and headings and sizes drawn from annotation priors. These sentences, along with the sentences derived from the ego agent's own perception, are not fused for feature enhancement but used only for text grounding, supervised by the corresponding footprints for $\mathcal{L}_{pos,i}$, the stated values for $\mathcal{L}_{val,i}$, and the stated attributes for $\mathcal{L}_{aux,i}$.

Hyperparameters. Table 5 lists the hyperparameter settings of GroundingV2X grouped by component. The first group gives the period bank of the Fourier number embedding in Section 3.3. The second group covers the routing and filtering of the text-to-feature enhancer in Section 3.4, including the number of selected cells K, the threshold ϵ, and the initialization of the learned temperature τ and invalid score b_0. The third and fourth groups give the temperature and near-duplicate threshold of numerical binding in Eq. (9) and the loss weights in Eq. (10). The last two groups specify the per-frame budgets of the sentences used in grounded pre-training and the probabilities of the simulated imperfections applied to the grounded text.

Table 5: Hyperparameters of GroundingV2X.

Group	Parameter	Value
Fourier number embedding	minimum period $T_{\min}$	0.08
	ratio ρ	$(1+\sqrt{5})/2$
	number of periods P	22
Text-to-feature enhancer	K for object sentences	4
	K for region sentences (regions vary in extent)	512
	threshold ϵ	10^{-3}
	temperature τ (init / range)	$e^{-2.66}$ / $[0.01, 1]$
	invalid score b_0 (init / min)	0.3 / 0.1
Numerical binding	temperature τ_v	0.07
	near-duplicate threshold δ	0.05
Loss weights	λ_{pos} / λ_{val} / λ_{aux}	0.5 / 0.2 / 0.5
Grounded pre-training	maximum received sentences per collaborator	40
	cross-task region sentences per frame	8
	virtual entities per frame	4
	received sentence dropout under cross-task coverage	0.4
	alignment-only sentences per frame (perceived / sampled)	24 / 30
Simulated imperfections	synonymous keywords probability	0.5
	permuted key-value order probability	0.5

B Additional Ablation Experiments

All ablations are conducted on the OPV2V dataset with $\mathrm{A}_{l,p}^{det}$ as the ego agent.

B.1 Effect of Synthetic Grounded Text and Pseudo-Labels

To evaluate the effect of the grounded pre-training with additional synthetic grounded text and pseudo-labels, we conduct an ablation study by removing the synthetic grounded text and pseudo-labels during pre-training and compare the performance under three groups of collaborators with increasing task heterogeneity, including the three detection collaborators of Table 1, the mixed group of one detection and two BEV segmentation collaborators of Table 3, and three BEV segmentation collaborators of the same type that send only region sentences. The results are reported in Table 6. As we can see, the model pre-trained without synthetic grounded text performs similarly to the full model when collaborating with collaborators in the same task. However, its performance drops significantly and becomes worse than accuracy without collaboration when receiving grounded text

from cross-task collaborators, indicating that unfamiliar sentences are misinterpreted and injected into the features as noise. In contrast, the full GroundingV2X retains the effective collaboration in task-heterogeneous groups, even when all collaborators are in different tasks. This demonstrates that synthetic grounded text is crucial for enabling cross-task collaboration.

Table 6: **Ablation study on the effect of synthetic grounded text and pseudo-labels during grounded pre-training.** We report AP@0.5 and AP@0.7 of the ego agent $\mathrm{A}_{l,p}^{det}$ with three groups of collaborators with increasing task heterogeneity, including three detection collaborators, the mixed group of one detection and two BEV segmentation collaborators, and three BEV segmentation collaborators, respectively.

Ego agent	$\mathrm{A}_{l,p}^{det}$					
Collaborators	$+\mathrm{A}_{c,r}^{det}$ $+\mathrm{A}_{l,s}^{det}$ $+\mathrm{A}_{c,e}^{det}$		$+\mathrm{A}_{c,r}^{det}$ $+\mathrm{A}_{l,s}^{bev}$ $+\mathrm{A}_{c,e}^{bev}$		$+\mathrm{A}_{l,s}^{bev}$ $+\mathrm{A}_{l,s}^{bev}$ $+\mathrm{A}_{l,s}^{bev}$	
Metric	AP@0.5	AP@0.7	AP@0.5	AP@0.7	AP@0.5	AP@0.7
No collaboration	0.748	0.606	0.748	0.606	0.748	0.606
w/o synthetic grounded text	0.901	0.813	0.555	0.512	0.484	0.446
GroundingV2X	0.901	0.813	0.849	0.744	0.843	0.738

B.2 Ablation Study on Numerical Encoding

We further evaluate the superiority of the proposed Fourier number embedding (FoNE) with irrational periods for numerical encoding of the grounded text. We compare it with the plain text encoding where each number is tokenized into ordinary subword tokens, the xVal (Golkar et al., 2023) encoding where the embedding of a single `[NUM]` token is scaled by its value, and the original FoNE with decimal periods (Zhou et al., 2026). The results of AP@0.5 and AP@0.7 with cumulatively increasing collaborators are reported in Table 7. As we can see, the FoNE with irrational periods achieves the strongest overall performance. Its advantage becomes clear as more collaborators are involved. The larger improvement under the stricter AP@0.7 criterion indicates that the irrational periods preserve numerical geometry more precisely when information from multiple messages is aggregated. Plain text performs surprisingly similarly to decimal-period FoNE in accuracy. However, tokenizing each numerical value into multiple subword tokens nearly doubles the number of tokens per sentence in our training data, substantially increasing the text encoder's activation memory and training cost. Overall, irrational FoNE achieves the best accuracy and efficiency by representing every value with a single numerically informed token and irrational period coefficients.

Table 7: **Ablation study on the numerical encoding methods.** AP@0.5 and AP@0.7 are reported as three heterogeneous detection collaborators are cumulatively added to ego agent $\mathrm{A}_{l,p}^{det}$. We compare the proposed Fourier number embedding (FoNE) with irrational periods to plain text, xVal, and decimal-period FoNE.

Numerical Encoding	AP@0.5↑			AP@0.7↑		
Ego agent $\mathrm{A}_{l,p}^{det}$	$+\mathrm{A}_{c,r}^{det}$	$+\mathrm{A}_{l,s}^{det}$	$+\mathrm{A}_{c,e}^{det}$	$+\mathrm{A}_{c,r}^{det}$	$+\mathrm{A}_{l,s}^{det}$	$+\mathrm{A}_{c,e}^{det}$
Plain text	0.842	0.889	0.889	0.740	0.785	0.786
xVal	0.831	0.845	0.845	0.748	0.753	0.753
FoNE (decimal periods)	0.844	0.890	0.890	0.739	0.785	0.785
FoNE (irrational periods, ours)	0.842	0.901	0.901	0.743	0.813	0.813

B.3 Robustness to Incomplete and Permuted Grounded Text

To evaluate the robustness of GroundingV2X to diverse grounded text structures and whether GroundingV2X grounds the received text by its specific key-value pairs rather than relying on a fixed sentence structure, we conduct an ablation study on the grounded text interface with randomly

permuted key-value order and dropped key-value pairs under different positional routing thresholds. The results are reported in Table 8. For full GroundingV2X, the accuracy of AP@0.5 and AP@0.7 slightly decreases with strict routing thresholds. Permuting the key-value pairs leaves performance virtually unchanged across all routing thresholds, indicating that GroundingV2X grounds the received text by their key-value pairs rather than relying on a fixed sentence structure. As the key-value pairs are removed at increasing ratios, the performance degrades gradually, and even after dropping 50% of them, the performance retains an AP@0.5 of 0.809 and an AP@0.7 of 0.703, outperforming no collaboration by 6.1 and 9.7 percentage points, respectively. This demonstrates that GroundingV2X is robust to incomplete grounded text. With the strict routing threshold $\epsilon = 0.999$, the performance after dropping 50% of the key-value pairs recovers to 0.826 of AP@0.5 and 0.735 of AP@0.7, indicating that stricter routing becomes advantageous under substantial information loss.

Table 8: **Ablation study on the robustness to incomplete and permuted grounded text.** AP@0.5 and AP@0.7 with three heterogeneous detection collaborators under different positional routing thresholds ϵ are reported.

Ego agent	$\mathrm{A}_{l,p}^{det}$					
Collaborators	$+\mathrm{A}_{c,r}^{det}$ $+\mathrm{A}_{l,s}^{det}$ $+\mathrm{A}_{c,e}^{det}$					
Routing Threshold	ϵ =0.001		ϵ=0.99		ϵ=0.999	
Conditions	AP@0.5	AP@0.7	AP@0.5	AP@0.7	AP@0.5	AP@0.7
No Collaboration	0.748	0.606	0.748	0.606	0.748	0.606
Full GroundingV2X	0.901	0.813	0.888	0.791	0.888	0.793
Permuted key-value order	0.901	0.814	0.888	0.791	0.888	0.792
10% of all key-value pairs dropped	0.870	0.771	0.861	0.758	0.868	0.770
30% of all key-value pairs dropped	0.826	0.720	0.826	0.721	0.839	0.744
50% of all key-value pairs dropped	0.809	0.703	0.813	0.716	0.826	0.735

B.4 Efficiency Analysis of Grounded Text Fusion

We further profile the actual latency of the grounded text fusion process with different numbers of received sentences. The latency is measured from the receipt of grounded text to the completion of text encoding and feature enhancement, using FP32 inference on an NVIDIA H100 GPU after a warm-up of 10 frames. As shown in Figure 5, collaboration with one to three collaborators follows a common approximately linear trend with respect to the total amount of sentences. The fitted slope is 0.94 ms per received sentence, with a fixed overhead of 16.4 ms, and the total latency virtually remains under 100 milliseconds even up to 70 sentences in one frame. This demonstrates that the grounded text fusion process is efficient and scalable, making it available for real-time V2X collaboration deployment.

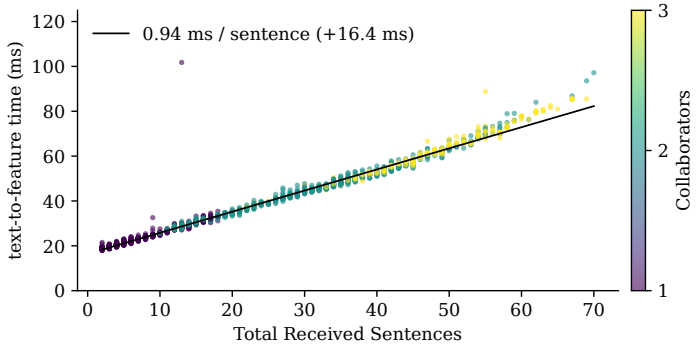


Figure 5: **Efficiency analysis of grounded text fusion.** We report the actual latency of the grounded text fusion process with different numbers of received sentences.